Leveraging Topological Structure

in Data Analysis, Machine Learning, and Visualization

Sourabh Palande

CMSE, Michigan State University

Seattle Children's Research Institute, March 24, 2023

Data is Shape, Shape is Data

Topology: Study of shape - How are things put together?

- Properties invariant under continuous deformations:
  - Translation, scaling, orientation, twisting, bending, etc..

TDA: Topological Data Analysis

- Collection of topological tools to:
  - Characterize and summarize the shape of data.
  - Main tools: Persistent Homology, Mapper
  - Utilize shape in data analysis, ML, visualization, etc.
- Applications:
  - Brain Networks,
  - Plant gene expression,
  - Scientific simulations,
  - ...

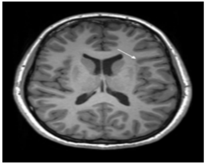
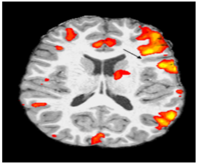
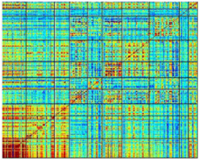
Overview

Part 1

Learning with Topological Features of Brain Networks

Learning with Brain Networks

Motivation: Leverage shape and structure of brain networks in ML

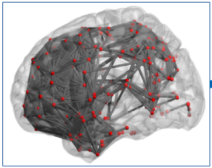
Idea: Brain Networks → Topological Features → Statistics / ML.

Contributions

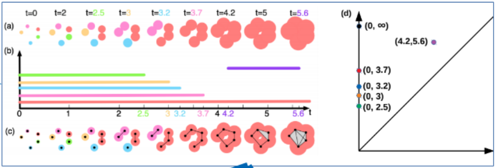
- Structural Networks: Statistical inference.
- Functional Networks: Regression (Predicting behavioral scores).
- Functional Networks: Classification (SVM, RF, neural nets).

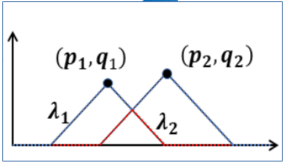
Topological Features

Brain Networks

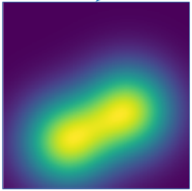


Persistent Homology





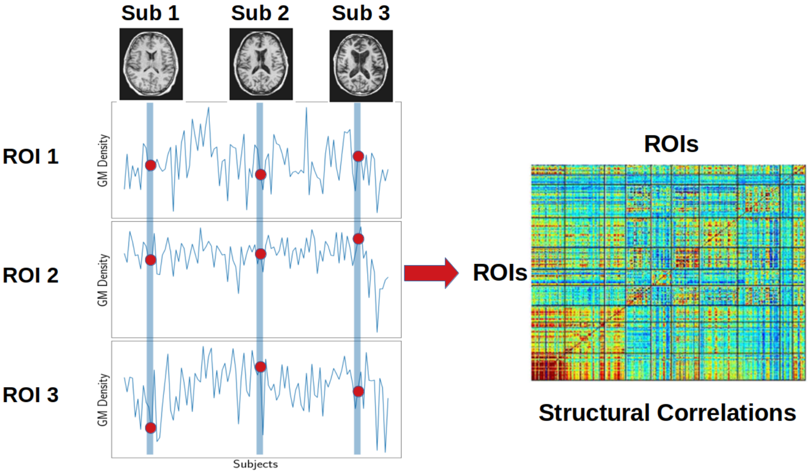
Persistence Landscapes



Persistence Images

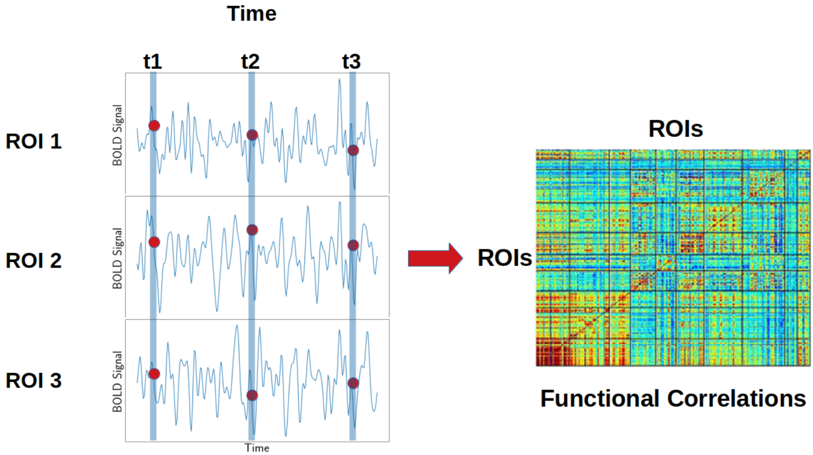
- **Kernels**
- **Projection layer for NN**

Structural Brain Networks



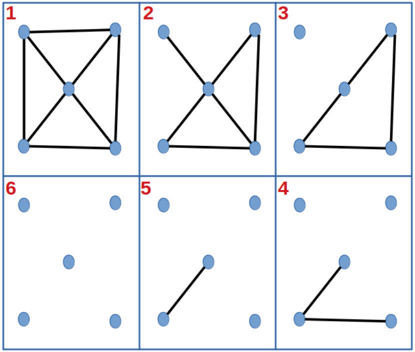
Encode shared structural influences across a group of subjects.

Functional Brain Networks



Encode level of synchronicity across time (for a single subject).

Graph Filtration



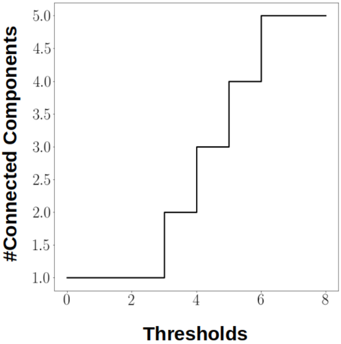


Figure: Graph filtration to compute β_0 (# connected components) curve.

Tracks changes in connectivity across a sequence of thresholds.

Persistent Homology

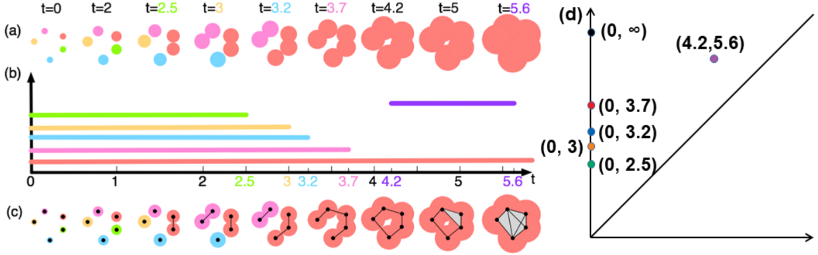
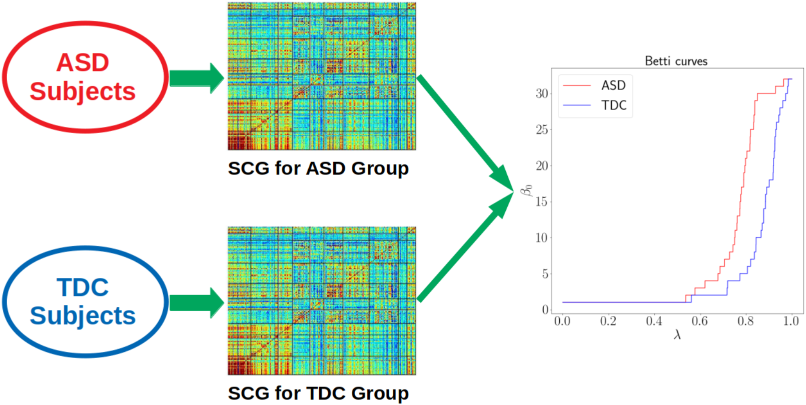


Figure: Persistent homology computation, represented as persistence barcodes in **(b)** and persistence diagrams (PDs) in **(d)**

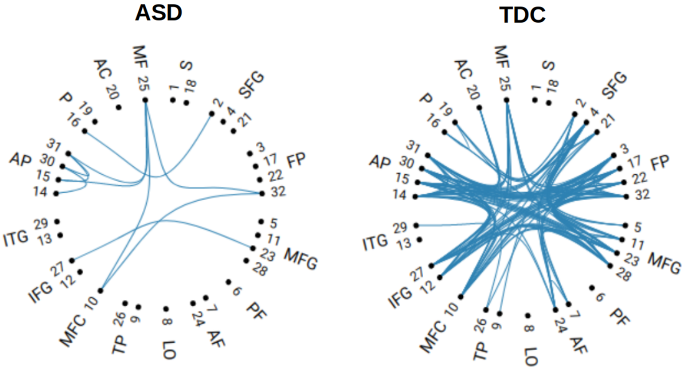
Tracks changes in topology across multiple scales

Statistical Inference with Structural Networks



- Permutation, Bootstrap tests
  - Test statistic: Largest gap between β_0 curves.

Statistical Inference with Structural Networks

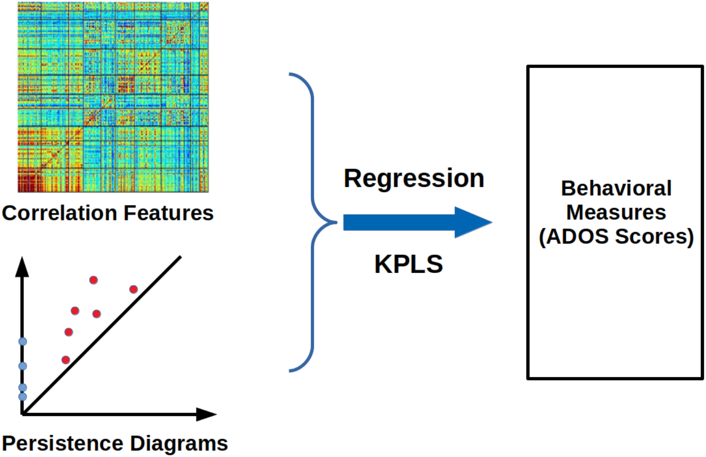


Main Result[1]**:** Evidence of abnormalities in gray matter regions associated with Salience Network.

---

[1] Palande, Jose, et al. 2019.

Relating Functional Networks to Behavioral Measures

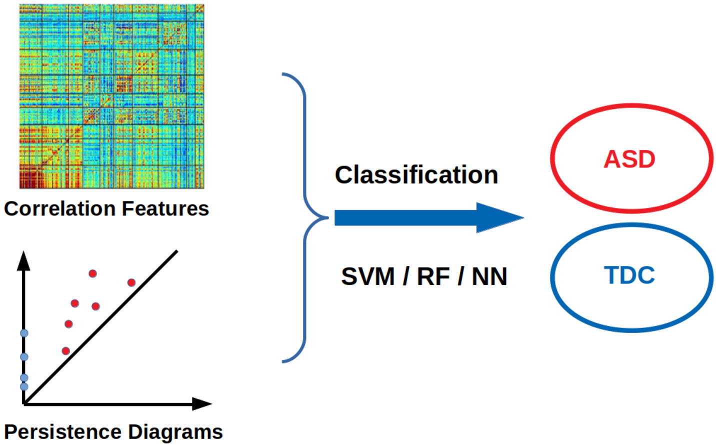


KPLS: Kernel Partial Least Squares Regression

Main Result[2]**:** Topological features improve predictive power.

[2] Wong et al. 2016.

Classification with Functional Networks



Main Result[3]**:** 69.9% classification accuracy.

[3] Rathore et al. 2019.

Main Results

- Regression[4]
  - Augmenting features through kernels (inner product matrices).
  - Adding topological features improves predictive power.
  - Only hybrid models provide statistically significant improvement.

- Classification[5]
  - Augmenting features through kernels (SVM, RF).
  - Custom layer for topological features (NN).
  - Hybrid models typically outperform.
  - Best accuracy: 69.9% (3-layer hybrid NN).
  - Issues due to data heterogeneity.

[4] Wong et al. 2016.

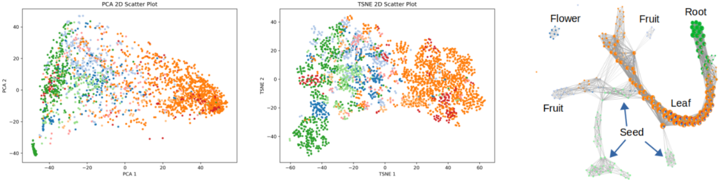
[5] Rathore et al. 2019.

Part 2

Visualizing the Shape of Gene Expression

Shape of Gene Expression

Motivation: Visual (meta-) analysis of gene expression across angiosperms



Idea: Apply Mapper to capture the shape of gene expression.

Contributions[6]**:**

- Interactive visualization built using Mapper.
- Hypotheses generation based on Mapper features.
- Identifying subsets of data and performing statistical analysis.

[6] Palande, Kaste, et al. 2022.

Mapper Algorithm

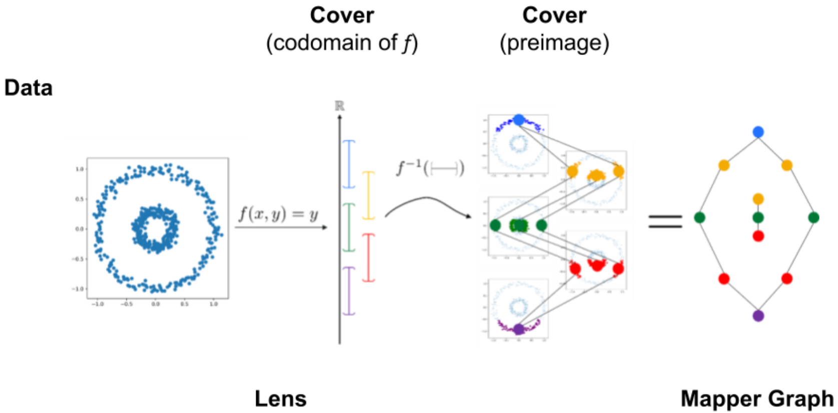


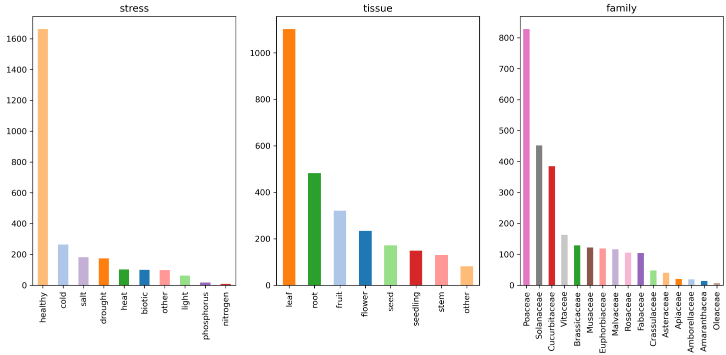
Figure: Mapper Algorithm

Mapper: Key Components

- Choice of lens: Domain / application dependent.
  - Only observe structure visible through specified lens.
  - Induce priors, domain knowledge.
- Choice of cover:
  - Determines connectivity, density of output graph.
  - Usually chosen by trial and error.
- Clustering algorithm:
  - Pick your favorite!
  - We stick to the default: DBSCAN.

Data

- 16 plant families, 54 distinct species.
- 8 tissue types, 9 biotic and abiotic stresses (+ healthy samples!)
- $\approx$ 3200 samples, 2671 left after processing.



Reducing Heterogeneity

- Cross-species analysis: Need correspondences!
- *Orthogroups:* Groups of homologous genes across species.
- TPM counts summed for genes in an orthogroup.
- Excluded multi-gene families with diverse functions.
- Excluded genes with high copy number.
- 2 million genes $\rightarrow$ 6328 orthogroups.
- Data combined into single expression matrix.
- 2671 Samples $\times$ 6328 orthogroups.

Dimension Reduction 1

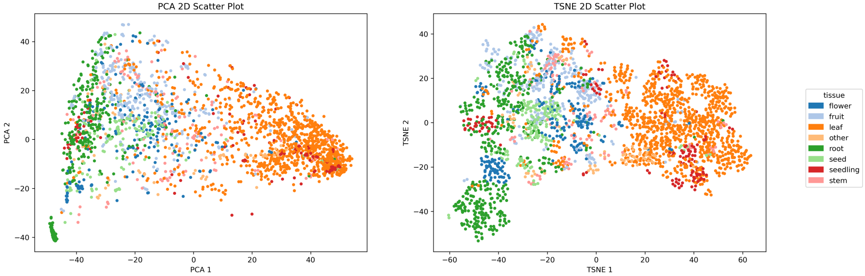


Figure: Dimension Reduction: Points colored by Tissue type.

Dimension Reduction 2

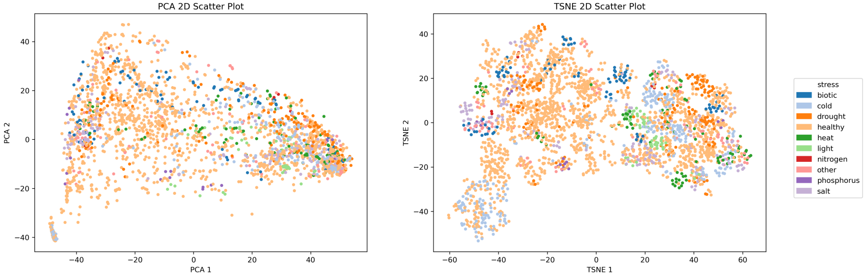


Figure: Dimension Reduction: Points colored by Stress type.

Creating Lenses[7]

- Two lenses: Tissue lens, Stress lens.
- Pick a base class:
  - *healthy* vs *stressed*, *leaf* vs *other*.
- Fit a linear model
  - *ideal* expression for base class.
- Project all samples on to the linear model.
- Residuals: Deviation from *ideal* expression.
- Use norm of the residual as lens.

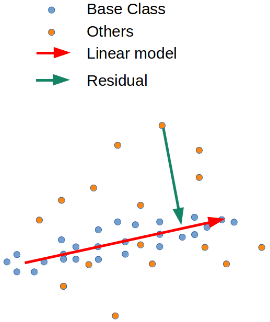


Figure: Creating lens

[7] Nicolau, Levine, and Carlsson 2011.

Lens Correlations and GO Enrichment

- Compute Lens-Orthogroup correlation.
- 2.5% most +ve (right tail).
- 2.5% most -ve (left tail).
- GO Enrichment Analysis for tail vs all.
- Use Arabidopsis genome as reference.
- GO Analysis tools:
  - https://pypi.org/project/goatools/

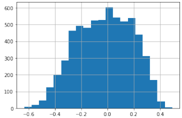


Figure: Leaf lens

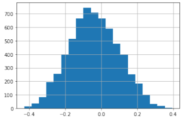


Figure: Stress lens

Go Enrichment Results

- Tissue lens: Captures photosynthetic vs non-photosynthetic divide.
- GO enrichment of +ve correlated orthogroups:
  - Core metabolic processes, development of non-photosynthetic tissues.
- GO enrichment of -ve correlated orthogroups:
  - Related to photosynthesis, response to light, chloroplast organization.
- Stress lens: healthy vs stressed gene expression
- GO enrichment of +ve correlated orthogroups:
  - Genes involved in stress response.
- GO enrichment of -ve correlated orthogroups:
  - Genes involved in growth and reproduction.

Mapper: Tissue Lens

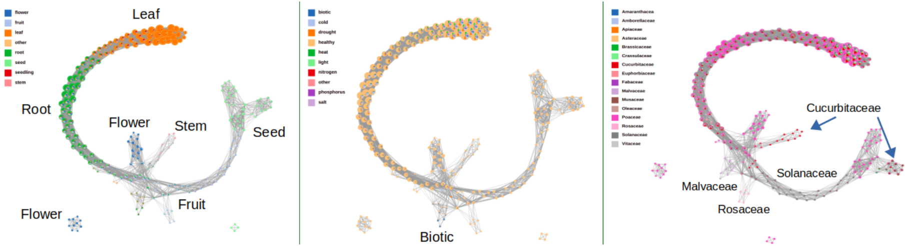


Figure: Tissue (leaf) Mapper Visualization

Mapper: Stress Lens

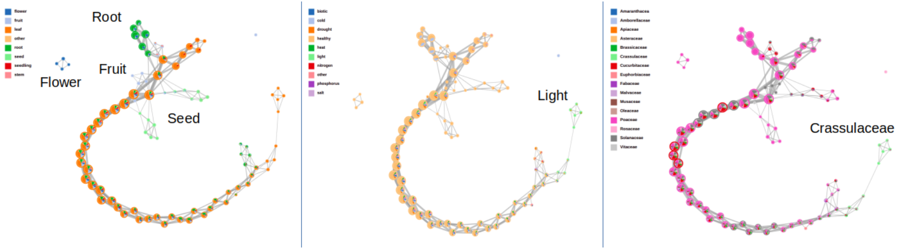
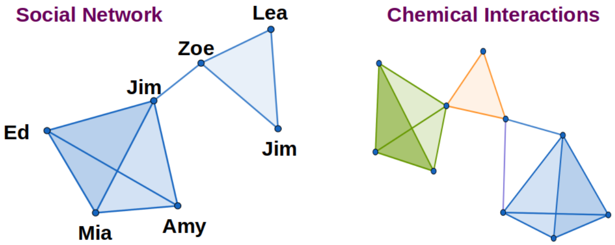


Figure: Stress Mapper Visualization

Part 3

Spectral Algorithms for Simplicial Complexes

Motivation: Leverage the topology of higher-order interactions in ML.



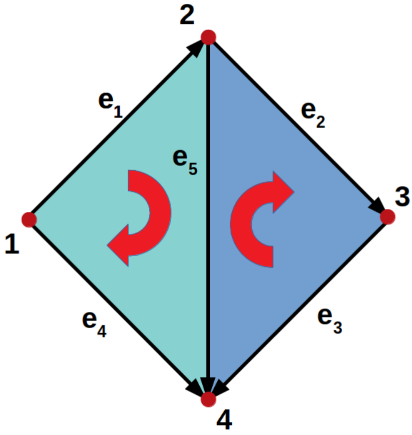
Idea: Operate directly on simplicial complexes.

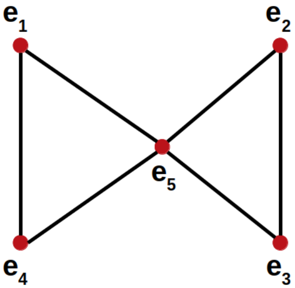
Contributions[8]**:**

- Label Propagation, Spectral Clustering for simplicial complexes.
- Spectral Sparsification.
- Random walks, Harmonics on simplicial complexes.

[8] Osting, Palande, and Wang 2020.

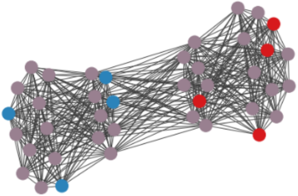
Dual Graph



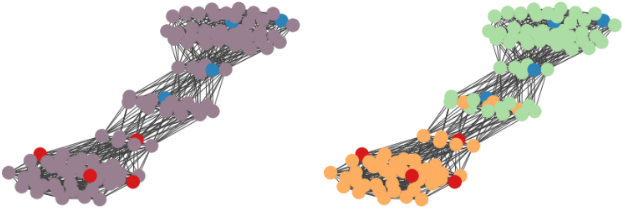


Label Propagation

Graphs



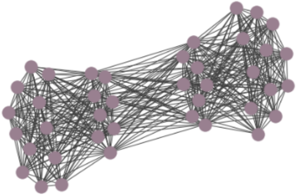
Simplicial Complexes[9]



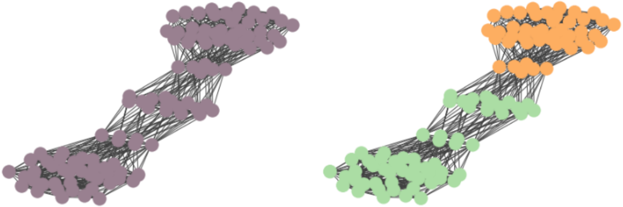
[9]We visualize the dual graph for simplicial complexes

Spectral Clustering

Graphs

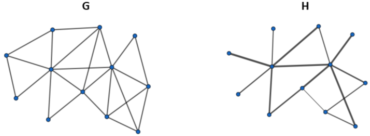


Simplicial Complexes



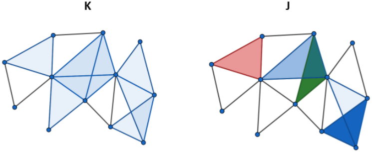
Sparsification: Preserving Spectral Properties

Graphs



$$(1-\epsilon)L_G \preceq \ L_H \preceq (1+\epsilon)L_G$$

Simplicial Complexes

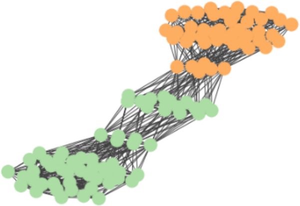


$$(1-\epsilon)\mathcal{L}_K \preceq \ \mathcal{L}_J \preceq (1+\epsilon)\mathcal{L}_K$$

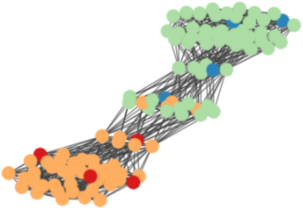
$$(1-\epsilon)x^T\mathcal{L}_K x \leq x^T\mathcal{L}_J x \leq (1+\epsilon)x^T\mathcal{L}_K x$$

Learning: Before and After Sparsification

Spectral Clustering

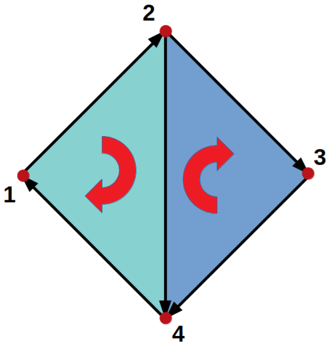


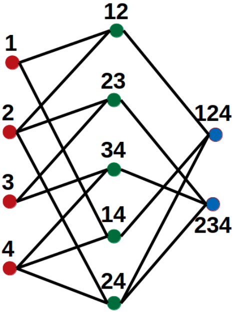
Label Propagation



Random Walk on Simplicial Complex

- We define random walk on the dual graph[10]
- Other versions have been explored in literature[11]
- We prove all are equivalent to random walk on the dual graph.





[10]Osting, Palande, and Wang 2020.

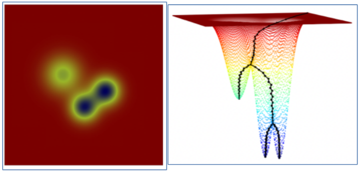
[11]Mukherjee and Steenbergen 2016; Parzanchevski and Rosenthal 2016.

Part 4

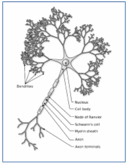
Aligning and Averaging Trees

Motivation: Perform computations on collections of trees.

Scalar Fields



Neurons



Trees

Idea: Optimal transport based alignment, combined with matrix sketching.

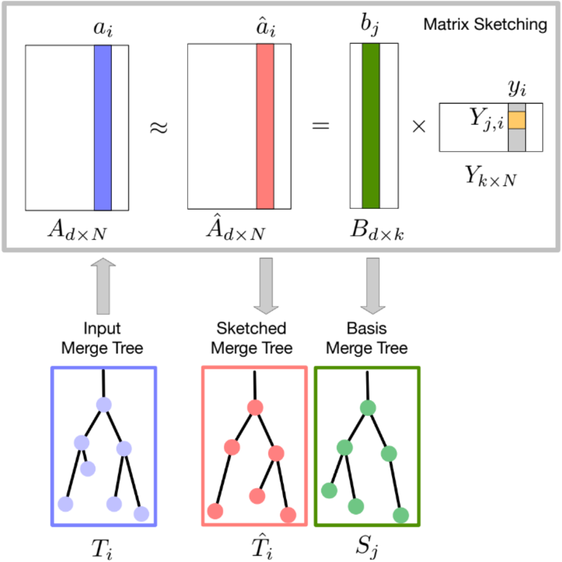
Contributions[12]**:**

- Adapt the Gromov-Wasserstein (GW) framework[13]
- Compute an average merge tree (Frechet mean)
- Compute a basis set of merge trees

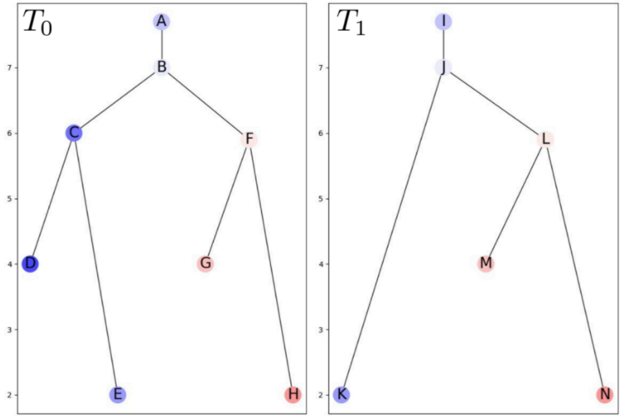
[12] Li, Palande, Yan, and Wang 2021.

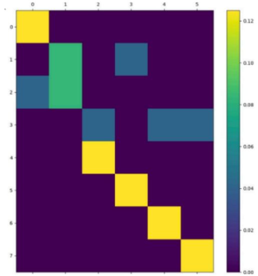
[13] Chowdhury and Needham 2019.

Matrix Sketching



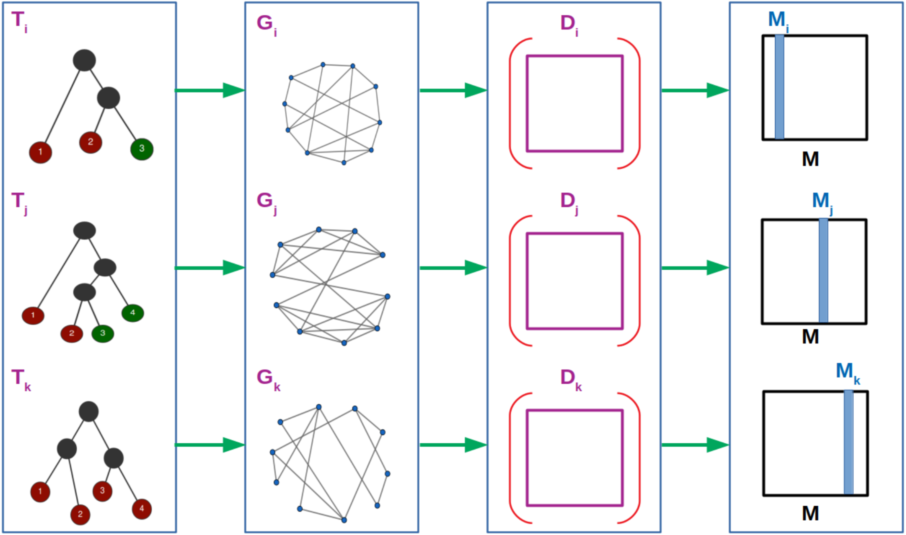
Tree Alignment



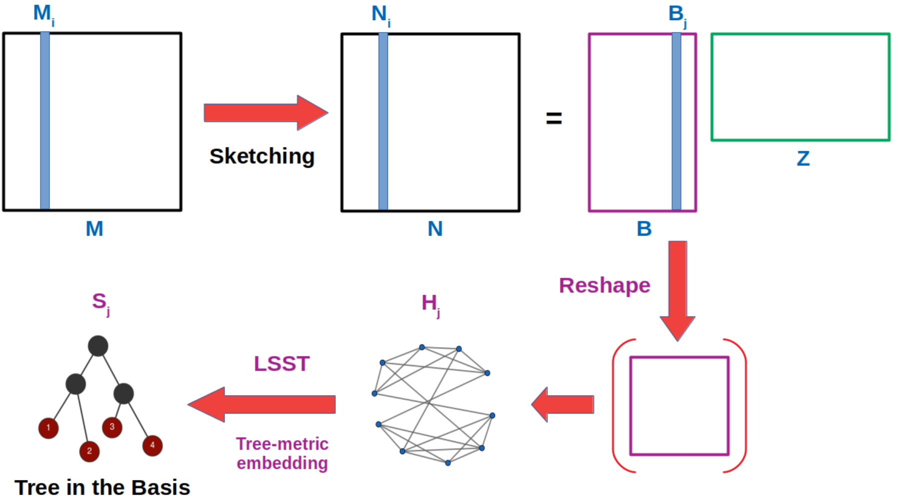


$$\begin{bmatrix} 0.125 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.083 & 0 & 0.042 & 0 & 0 \\ 0.042 & 0.083 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.042 & 0 & 0.042 & 0.042 \\ 0 & 0 & 0.125 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.125 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.125 \end{bmatrix}$$

Gromov-Wasserstein Mapping



Tree Sketching Pipeline



Merge Tree

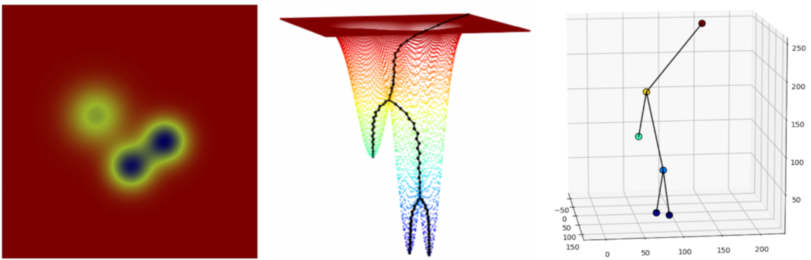
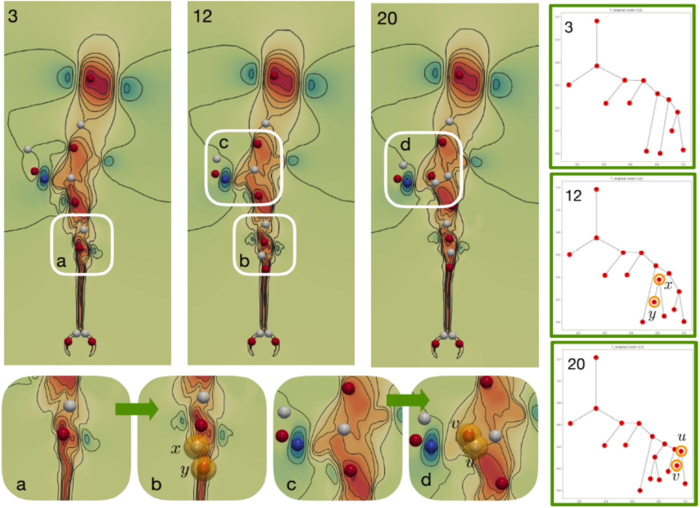
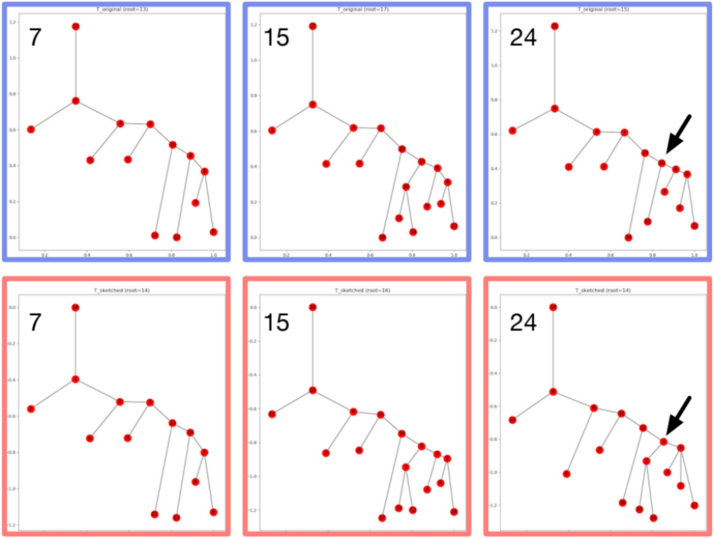


Figure: Merge tree from a scalar field [LinWangMunch2020]

Application: Heated Cylinder Simulation



Tree Sketching Pipeline



Part 5

Recap and Future Directions

Recap

Leveraging topology in data analysis, ML, and visualization.

- Feature engineering: Brain network applications.
  - Statistical Inference.
  - Regression.
  - Classification.
- ML Algorithms: Learning on Simplicial Complexes.
- Dimension Reduction: Tree alignment and sketching.
- Visualization: plant gene expression.

Ongoing Work

Leveraging topology in data analysis, ML, and visualization.

- Improving Mapper.
  - Systematic parameter tuning.
  - Fuzzy Clustering, mixture models for cover parameter.
  - Learning lens function through topological optimization.
- Evaluating Arabidopsis as model species.
  - Training ML models on Arabidopsis gene expression.
    - Using full gene set 37K.
    - Using 2671 orthogroup reference genes.
  - Tissue classification accuracy:
    - Arabidopsis: 98%
    - Angiosperms: 64%
  - Is Arabidopsis a good model?

Future Direction

Proposal: Hypergraph models and methods for *omics.

- Genome-wide hypergraph construction.
  - Graph Coarsening.
  - Mapper / Fuzzy clustering.
- Machine learning on hypergraphs.
  - Extending graph ML to hypergraphs.
  - Stochastic processes / dynamical systems on hypergraphs.
  - Physics inspired / Physics based ML models.
- Hypergraph alignment. (Optimal transport!)
- Trained model adaptation through alignment.
- Cross-specie / multi-specie ML models.

Part 6

References

References I

[CN19] Samir Chowdhury and Tom Needham. "Gromov-wasserstein averaging in a riemannian framework". preprint, arXiv:1910.04308. 2019. arXiv: 1910.04308 [math.MG].

[Li+21] Mingzhe Li, Sourabh Palande, Lin Yan, and Bei Wang. "Sketching merge trees for scientific data visualization". arXiv:2101.03196 [cs.CG]. 2021.

[MS16] Sayan Mukherjee and John Steenbergen. "Random walks on simplicial complexes and harmonics". In: *Random Structures & Algorithms* 49.2 (2016), pp. 379–405. DOI: 10.1002/rsa.20645.

[NLC11] Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival". In: *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 7265–7270. DOI: 10.1073/pnas.1102826108.

References II

[OPW20] Braxton Osting, Sourabh Palande, and Bei Wang. “Spectral sparsification of simplicial complexes for clustering and label propagation.”. Journal of Computational Geometry (JoCG), to appear. 2020.

[Pal+19] Sourabh Palande, Vipin Jose, et al. “Revisiting abnormalities in brain network architecture underlying autism using topology-inspired statistical inference.”. In: *Brain Connectivity* 9.1 (2019), pp. 13–21.

[Pal+22] Sourabh Palande, Joshua A.M. Kaste, et al. “The topological shape of gene expression across the evolution of flowering plants”. bioRxiv:2022.09.07.506951. 2022. DOI: 10.1101/2022.09.07.506951.

[PR16] Ori Parzanchevski and Ron Rosenthal. “Simplicial complexes: Spectrum, homology and random walks”. In: *Random Structures & Algorithms* 50.2 (2016), pp. 225–261. DOI: 10.1002/rsa.20657.

References III

[Rat+19] Archit Rathore et al. "Autism classification using topological features and deep learning: a cautionary tale.". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer International Publishing, 2019, pp. 736–744.

[Won+16] Eleanor Wong et al. "Kernel partial least squares regression for relating functional brain network topology to clinical measures of behavior". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 1303–1306. DOI: 10.1109/isbi.2016.7493506.

Part 7

Extra Slides